



**School of Informatics, University of Edinburgh**

---

**Institute of Perception, Action and Behaviour**

**Active Perception in Navigation of Partially Observable Grid Worlds**

by

Paul Crook, Gillian Hayes

**Informatics Research Report EDI-INF-RR-0181**

---

**School of Informatics**  
<http://www.informatics.ed.ac.uk/>

**September 2003**

# Active Perception in Navigation of Partially Observable Grid Worlds

Paul Crook, Gillian Hayes

Informatics Research Report EDI-INF-RR-0181

SCHOOL *of* INFORMATICS

Institute of Perception, Action and Behaviour

September 2003

In proceeding of the Sixth European Workshop on Reinforcement Learning (EWRL-6, 2003)

## **Abstract :**

Research into using reinforcement learning to find optimal solutions to tasks where only partial information is available, i.e. partially observable Markovian decision processes (POMDPs), has traditionally focused on augmenting learning algorithms with memory or the ability to build internal models of the world. Our approach differs in that we consider agents with active perception, i.e. agents who exercise control over the sensory input they obtain from the world. Our conjecture is that agents should be able to learn to use active perception to find optimal solutions to what are otherwise POMDPs; and further, that this is possible using basic reinforcement learning techniques that do not employ memory or internal models.

This two page paper presents our preliminary empirical investigation into this conjecture. We present some suggestive results, though at this stage our work is a limited proof of concept, focusing on a single grid world problem with no comparison against the efficacy of reinforcement learning techniques enhanced with memory or modelling abilities.

**Keywords :** reinforcement learning, active perception, partially observable Markovian decision processes, POMDPs

Copyright © 2003 by The University of Edinburgh. All Rights Reserved

The authors and the University of Edinburgh retain the right to reproduce and publish this paper for non-commercial purposes.

Permission is granted for this report to be reproduced by others for non-commercial purposes as long as this copyright notice is reprinted in full in any reproduction. Applications to make other use of the material should be addressed in the first instance to Copyright Permissions, School of Informatics, The University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, Scotland.

---

# Active Perception in Navigation of Partially Observable Grid Worlds

---

Paul A. Crook  
Gillian Hayes

PAULC@DAI.ED.AC.UK  
GMH@INF.ED.AC.UK

Institute of Perception Action and Behaviour (IPAB), School of Informatics, University of Edinburgh,  
5 Forrest Hill, Edinburgh, EH8 9PR, UK

## 1. Introduction

Research into using reinforcement learning to find optimal solutions to tasks where only partial information is available, *i.e.* partially observable Markovian decision processes (POMDPs), has traditionally focused on augmenting learning algorithms with memory or the ability to build internal models of the world. Our approach differs in that we consider agents with *active perception*, *i.e.* agents who exercise control over the sensory input they obtain from the world. Our conjecture is that agents should be able to learn to use active perception to find optimal solutions to what are otherwise POMDPs; and further, that this is possible using basic reinforcement learning techniques that do not employ memory or internal models.

This abstract presents our preliminary empirical investigation into this conjecture. We present some suggestive results, though at this stage our work is a limited proof of concept, focusing on a single grid world problem with no comparison against the efficacy of reinforcement learning techniques enhanced with memory or modelling abilities.

## 2. Background

Whitehead, (1992, chp.5) identifies two distinct problems that occur when using reinforcement learning to find solutions to POMDPs: *local* and *global impairment*. An example of local impairment is a robot stood at one of two identical looking T-junctions, at one it ought to turn left, at the other turn right. As it is unable to distinguish between the two states it regards them as one and the same, we use the phrase *aliased state* to refer to such states. Basic reinforcement learning associates a single action with a given state and a single value estimate with each state (or state-action pair). There are therefore two problems that occur locally: (i) the single action that the robot learns to execute will, at one of the T-junctions, be inconsistent with the true underlying state; (ii) averaging will occur of what should be independent value estimates. The latter effect is independent of whether the action to be executed is inconsistent or not. Global impairment oc-

curs because 1-step backup, etc. will propagate these inconsistent state values when updating the state values of otherwise consistent states. This can spread to affect the whole of the learnt policy.

Loch and Singh (1998) showed that the use of eligibility traces allows reinforcement learning to find optimal memoryless policies (see table 1 for definition) in partially observable grid world navigation problems. The use of eligibility traces succeeds where 1-step backup reinforcement learning algorithms fail (Whitehead, 1992), by partially correcting the state valuation errors caused by global impairment.

Our conjecture arises as follows. If an agent with active perception can learn to find an unambiguous reference when it encounters an aliased state, it can then select the optimal action for that state, thus dealing with one of the effects of local impairment. Provided that eligibility traces sufficiently correct for global impairment, the agent should then be able learn optimal solutions despite the presence of aliased states.

## 3. Experiments

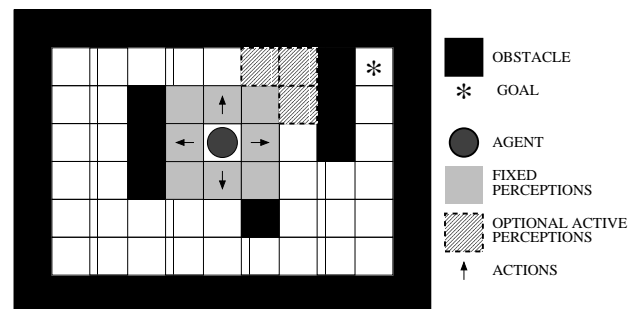


Figure 1. Sutton's grid world and example agent.

We use Sutton's grid world (figure 1) as modified by Littman (1994). An agent starts in any grid square not containing an obstacle and has to reach the goal square indicated by an asterisk. The agent can choose between four *physical actions*: move north, south, east or west. It receives a reward of  $-1$  for any action that does not move it towards the goal and a reward of  $0$  for reaching the goal state. If the agent attempts to move

towards an obstacle it remains in its current location although still receiving a reward of  $-1$ .

We compare two agents. One agent’s state representation is formed by observing the eight squares adjacent to its current location. This agent experiences state aliasing in multiple locations of Sutton’s grid world (see Littman (1994) for details). The second agent has the same basic state input as the first, but has eight additional *perceptual actions* that it can select: look north, north east, east, south east, south, south west, west or north west. On selecting one of these perceptual actions its input state is increased to include information from three additional squares in the direction of it choosing (e.g. the three hatched squares in figure 1). It receives a reward of  $-1$  for selecting a perceptual action. Its input state reverts to the eight adjacent squares after selecting a physical action. The performance of the first agent, who has no active perception, forms a baseline against which we compare the performance of the second.

The reinforcement learning algorithm SARSA( $\lambda$ ) with replacement traces (Sutton and Barto, (1998, p181)) was used with both agents. Parameters used: discount rate  $\gamma = 0.9$ ; probability of exploratory action  $\epsilon$  initially 0.2, decaying linearly to zero by the 500,000<sup>th</sup> action-learning step; learning rates  $\alpha$  0.02 and 0.05; eligibility trace decay rates  $\lambda$  0.9 and 0.99. In an ad-hoc search of the parameter space, these values gave rise to the minimum mean total of perceptual actions and the maximum number of physically optimal solutions for the second agent. One hundred repetitions were run for each set of parameters. Convergence of policies in each case was judged (by eye) to have occurred after 500,000 action-learning steps.

#### 4. Results, Discussion & Future Work

<b>Physically Optimal</b>	Reaches goal in the minimum total number of <i>physical</i> actions from all start states	404 physical actions
<b>Better than Optimal Memoryless</b>	A lower total of physical actions than can be achieved by a memoryless agent	>404 <416
<b>Optimal Memoryless</b>	The minimum total of physical actions that can be achieved by a memoryless agent	416 physical actions
<b>Satisficing</b>	<b>Reaches goal from all start states</b> (in our experiments a run from a given start state is terminated and counted as unsuccessful if 1000 action steps are exceeded)	

Table 1. Classification of policies derived from Littman (1994)’s definitions. Values given for Sutton’s grid world

The policies obtained are classified in accordance with the criteria in table 1. As can be seen from figure 2 the second agent can indeed learn physically optimal policies. Further, it can find policies whose mean total physical actions are significantly less (407.4) than the

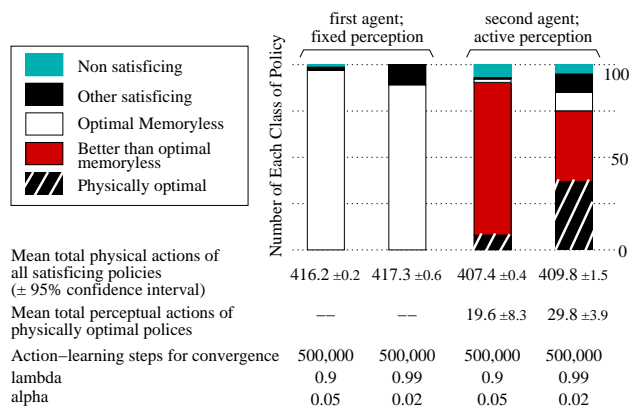


Figure 2. Summary of results

best achievable by a memoryless agent (416). However, the second agent rarely learns totally optimal policies; *i.e.* optimal number of physical and perceptual actions. Most physically optimal policies use more perceptual actions than is strictly necessary. We believe a totally optimal policy to require a total of 12 perceptual and 404 physical actions. In selecting the parameters used, we found that increasing the period over which exploratory actions can occur, *i.e.* decreasing the rate of decay of  $\epsilon$ , significantly reduces the number of redundant perceptual actions in each policy. Unexpectedly, increasing exploration has only a small effect on the number of physically optimal solutions found. The number of physically optimal solutions appears to be dependent on the values of  $\alpha$  and  $\lambda$ .

For both agents not all the policies converge to satisficing solutions. What impediment prevents convergence, why more physically optimal solutions are not found, whether these results can generalise to other grid world problems, and comparisons with other approaches are all areas for future research.

#### References

- Littman, M. L. (1994). Memoryless policies: Theoretical limitations and practical results. *From Animals to Animals 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior (SAB’94)* (pp. 238–245). The MIT Press, Cambridge, MA.
- Loch, J., & Singh, S. (1998). Using eligibility traces to find the best memoryless policy in partially observable Markov decision processes. *Proceedings of the Fifteenth International Conference on Machine Learning (ICML’98)* (pp. 323–331). Morgan Kaufmann, San Francisco, CA.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: The MIT Press.
- Whitehead, S. D. (1992). *Reinforcement learning for the adaptive control of perception and action*. Doctoral dissertation, University of Rochester, Department of Computer Science, Rochester, New York.